EDK | CDIP | CDPE | CDEP |
Schweizerische Konferenz der kantonalen Erziehungsdirektoren
Conférence suisse des directeurs cantonaux de l'instruction publique
Conferenza svizzera dei direttori cantonali della pubblica educazione
Conferenza svizra dals directurs chantunals da l'educaziun publica

HARMOS - Développement de standards de formation

Rapport final du groupe Méthodologie

Erich Ramseier
Abteilung Bildungsplanung und Evaluation der Erziehungsdirektion des Kantons Bern

Urs Moser
Institut für Bildungsevaluation, Université de Zurich

Jean Moreau
Unité de recherche pour le pilotage des systèmes pédagogiques du canton de Vaud

Jean-Philippe Antonietti
Institut de mathématiques appliquées, Université de Lausanne

Berne, juillet 2008

Table des matières

La Conférence suisse des directrices et directeurs cantonaux de l'instruction publique (CDIP) a confié à un groupe de projet scientifique le mandat d'élaborer des standards de base pour les quatre disciplines suivantes: langue de scolarisation, mathématiques, langues étrangères, sciences expérimentales. Les standards de base doivent se fonder sur un modèle de compétence. Ce travail de développement comprend également la validation empirique de ces modèles de compétence (étude de validation).

Dans ce travail de développement, la responsabilité scientifique pour chaque discipline a été confiée à un consortium. A cela s'est ajouté un groupe Méthodologie[1], constitué par le mandat du 1er avril 2005, qui, pendant tout le déroulement du projet, a accompagné les consortiums pour les aspects méthodologiques et techniques pouvant influencer les modèles de compétence des consortiums. Les tâches du groupe comportaient notamment la conception de l'étude de validation et l'évaluation des données en perspective de l'étalonnage et de la validation des modèles de compétence. Le groupe devait également rédiger un rapport final présentant un bilan des différentes procédures et des difficultés méthodologiques, et formulant des recommandations pour les étapes futures.

Après une brève description des tâches confiées au groupe, le présent rapport indique comment évaluer les résultats et quelles conclusions en tirer. Il ne documente pas chaque recherche, raison pour laquelle il ne constitue pas une documentation scientifique de l'étude de validation.

1. Mode de travail

Organisation interne

Le groupe Méthodologie avec Erich Ramseier comme coordinateur, s'est réuni cinq à six fois par an entre 2005 et 2007, et deux fois en 2008. En plus de ces réunions, il y a eu de nombreuses rencontres avec les consortiums et des contacts bilatéraux. Chacun des membres était responsable de deux thématiques pour l'étalonnage, comme le présente le tableau ci-dessous:

	6e année	*9e année*
langue de scolarisation	Moser	Moreau
mathématiques	Antonietti	Ramseier
sciences expérimentales	Moreau	Ramseier
	lecture, C-Test	*Écriture, compréhension orale*
langues étrangères	Antonietti	Moser

En outre, Urs Moser, Erich Ramseier et Jean Moreau ont assuré le suivi de l'étalonnage des études pilotes effectuées en 2e année sur la langue de scolarisation, les mathématiques ou les sciences expérimentales.

[1] Jean-Philippe Antonietti, Université de Lausanne; Jean Moreau, URSP Lausanne; Urs Moser, IBE Université de Zurich; Erich Ramseier, direction de l'instruction publique du canton de Berne

Groupe stratégique

La fonction du groupe Méthodologie était double; d'une part l'activité de conseil et de réalisation et, d'autre part, l'élaboration de directives contraignantes pour la réalisation scientifique de l'étude de validation. Composé de représentants de la direction des différents consortiums et du groupe Méthodologie, un groupe stratégique, sous la houlette d'Olivier Maradan, a été mis sur pied pour prendre des décisions contraignantes (secrétariat: Max Mangold). Les décisions communes ne portaient pas seulement sur des questions méthodologiques, mais également sur les traits fondamentaux du modèle de compétence, la terminologie commune et le type de produits (cf. guide). Au sein du groupe stratégique, des thématiques méthodologiques spécifiques ont été présentées, par exemple sur le Standard-Setting (formation continue).

Evaluation: le groupe stratégique a été certes créé alors que le projet était déjà bien avancé, mais il en a été un instrument central. Il a veillé à assurer une base commune minimale entre les consortiums; un calendrier très serré a toutefois contrecarré un élargissement de cette base commune.

Contacts avec les experts externes

Ayant reçu l'autorisation de faire appel à des experts externes, le groupe Méthodologie a rencontré, en été 2005, Ray Adams de l'Université de Melbourne/responsable de PISA, pour tirer au clair la conception du test de validation et les problèmes d'étalonnage. Pendant l'étalonnage, Ray Adams a été sollicité plusieurs fois pour donner son avis par courriel.

Dans le cadre de la convention DACHL, Urs Moser et Erich Ramseier se sont rendus à deux reprises à un colloque et un échange d'expériences avec leurs homologues d'Allemagne, d'Autriche et du Luxembourg qui se trouvent dans des conditions bien mieux définies de réalisation de tâches semblables (cf. IQB Berlin).

Evaluation: la poursuite et le renforcement des contacts internationaux au niveau de la méthodologie de la mesure des compétences s'avèrent indispensables. En effet, la Suisse, et tout particulièrement la Suisse alémanique, ne peut pas s'appuyer sur une tradition bien établie. Les discussions avec des chercheurs provenant des pays voisins ont mis en évidence que ces derniers engagent bien davantage de ressources dans le développement des modèles de compétence et la validation empirique. De plus, à la différence de la Suisse, ces pays ont instauré des types d'organisation permanents pour le développement de tests de résultats.

2. Guide

Le groupe Méthodologie a rédigé un guide descriptif des exigences posées aux modèles de compétence et à l'étude de validation. Il donne une bonne vue d'ensemble du projet en dépit de l'aspect provisoire de la conception, de l'échantillon et de la réalisation du test puisqu'il ne s'agit que du stade précédant la planification définitive. Signalons ici que les enquêtes sur les langues étrangères et les sciences expérimentales ont été avancées de 2008 (calendrier du guide) au printemps 2007 pour être réalisées en même temps que les enquêtes sur les mathématiques et la langue première.

Dans la partie théorique, le guide contient les décisions prises par le groupe stratégique concernant le modèle de compétence et, dans la partie principale, il décrit les exigences en matière d'élaboration des exercices et de préparation du test.

Evaluation: les indications concernant la préparation des tests et la formulation des exercices ont été respectées par les consortiums ce qui a permis, en dépit de la différence culturelle dans les diverses disciplines, de maintenir un bon niveau de communication avec le groupe Méthodologie et d'aboutir à un test commun susceptible d'être évalué.

3. Prétest

Une fois développés et testés à l'échelle réduite ou dans ce qu'on appelle des laboratoires cognitifs, tous les exercices sont habituellement soumis à un prétest au moins avant leur utilisation sur le terrain. Le calendrier très serré d'HarmoS ne l'a pas permis. A la demande pressante du groupe Méthodologie, les consortiums ont effectué, en 2006, des prétests avec un nombre limité d'exercices. D'une part, les prétests ont fourni des informations sur les exercices pris en considération; d'autre part, ils ont mis en lumière les différentes étapes à franchir lors d'une mesure empirique des résultats et illustré les résultats à attendre de l'étalonnage et la manière de les gérer. Le prétest a de ce fait rempli une fonction essentielle de formation continue.

Evaluation: l'expérimentation incomplète des exercices dans les prétests doit être considéré comme une procédure d'urgence et ne doit donc pas être reproduite dans les tests ultérieurs sur le monitorage de l'éducation. Il faudrait notamment détecter les difficultés liées à la traduction à ce moment-là et non lors de la validation proprement dite. D'une manière générale, la charge représentée par le développement de tests en trois langues a été sous-estimée dans le projet HarmoS.

Il n'en reste pas moins que l'expérimentation de la procédure empirique et l'interaction entre consortiums et groupe Méthodologie se sont révélées très précieuses; l'absence du prétest et, par là-même de la formation continue, aurait mis en péril le succès de l'étude de validation.

4. Conception de l'étude de validation

Le groupe stratégique a retenu qu'en raison de la prééminence de l'exhaustivité du contenu, les modèles de compétence ne doivent pas se limiter aux éléments que l'étude de validation permet déjà de vérifier empiriquement. Dans ces conditions et compte tenu entre autres choses de la contrainte du temps, l'étude de validation représentative a été limitée aux 6e et 9e année; en outre, selon la discipline, certains domaines ou éléments de compétence n'ont pas été inclus dans l'étude de validation avec les exercices qui allaient de pair. En même temps, des études réduites ont été effectuées en 2e année pour la langue première, les mathématiques et les sciences expérimentales, avec des échantillons non représentatifs. Ceci est parfaitement suffisant pour le classement des exercices tant que l'échantillon englobe tout l'éventail des compétences individuelles.

Objectif

L'étude de validation sert d'abord à vérifier les modèles de compétence (1er objectif). Pour ce faire, la description de la compétence doit être étayée empiriquement par les domaines, aspects et niveaux. En amont, il faut donc qu'il y ait un nombre suffisant d'exercices à tester et que leur degré de difficulté soit connu pour saisir les combinaisons de domaines, aspects et niveaux et les illustrer par des exercices types. A cela s'ajoute que l'étude soit représentative de la distribution des élèves sur les niveaux de compétence (2e objectif). On indiquera notamment quelle proportion de la population scolaire remplit aujourd'hui les standards de base proposés.

Echantillon et pondération

Pour remplir le premier objectif de l'étude, il faut inclure autant d'élèves que nécessaire à une estimation suffisamment exacte du degré de difficulté des exercices dans les différentes régions linguistiques. Le guide esquisse à la p. 17 ss. des réflexions qui vont dans ce sens. En tenant compte des défections prévisibles et à la condition que les personnes testées passent deux demi-journées de test, on peut raisonnablement prévoir un échantillon brut de 6.600 élèves par année testée. Le deuxième objectif exige la représentativité de l'échantillon. Renaud (2007) décrit comment s'est constitué un échantillon en deux étapes (choix des écoles et sélection de deux années).

Evaluation: la composition de l'échantillon a fait ses preuves. Les données ainsi obtenues permettent de décrire le degré de difficulté des exercices et la distribution des compétences avec l'exactitude requise.

Conception des tests

De nombreux exercices et leur contexte ont fait l'objet d'un examen approfondi dans les tests HarmoS pour vérifier s'ils étaient appropriés pour caractériser un modèle de compétence. Cela implique de nombreux exercices alors qu'une personne testée ne peut en traiter que quelques uns pendant le temps imparti. Il

faut donc trouver l'attribution appropriée des exercices et des moments de tests pour rendre possible leur évaluation optimale (matrix sampling).

Les quatre consortiums ont développé les exercices à traiter (davantage d'informations dans le guide p. 10 ss.) et les ont regroupés en une grappe (cluster), qui peut être traitée dans un laps de temps de 20 à 30 minutes. En tout, cela représente 209 grappes, qui diffèrent par le contenu et qui existent en deux ou trois langues. Un procédé spécial a permis d'imprimer chaque grappe dans un petit cahier de tests séparé et d'attribuer individuellement chaque cahier à une personne testée en indiquant à quel moment ce cahier devait être traité pendant la réalisation du test.

La conception utilisée pour le test a bénéficié de cette possibilité et a contribué, en tenant compte des exigences des consortiums, à une dispersion des exercices la plus large et la plus équilibrée possible. Tous les détails de la procédure figurent en annexe. Dans l'ensemble, environ 100.000 liens individuels ont été établis entre grappes d'exercices, personnes testées et moments de tests; seules quelques rares personnes testées se sont attaquées avec succès à la même combinaison d'exercices.

Evaluation: l'inclusion d'un procédé professionnel d'impression (centre d'impression de la NZZ) a permis l'affectation individuelle des grappes, qui est très coûteuse, aux élèves choisis au hasard. La mise en œuvre de cette conception de test extrêmement complexe a permis une répartition équilibrée des exercices entre les élèves, ce qui a été bénéfique pour l'étalonnage consécutif. Cet équilibre ne vaut qu'entre les grappes. L'influence de la séquence des exercices au sein d'une grappe reste incontrôlable (comme à l'accoutumée).

Questionnaire pour les élèves

Un bref questionnaire a été élaboré pour connaître l'origine culturelle et sociale, l'intérêt pour telle ou telle discipline, tout particulièrement dans les domaines des mathématiques et des sciences expérimentales, pour compléter les tests de performance et les informations extraites des listes de contrôle des élèves.

Evaluation: il est indispensable pour la validation des exercices du test de saisir des caractéristiques essentielles qui sont en lien direct avec les résultats scolaires. Ces dernières ont été prises en compte dans l'étalonnage. Jusqu'où doit aller l'évaluation de l'intérêt personnel; à ce propose, il règne une part d'incertitude. Le fait d'être issu de la migration n'a pas non plus été évalué complètement.

5. Réalisation de l'étude de validation

Réalisation du test

Conformément à la planification concertée entre le Secrétariat général de la CDIP et le groupe Méthodologie, la saisie des données, la traduction et la production des cahiers de tests ont été prises en main par le Secrétariat général de la CDIP. Ces tâches comprenaient l'identification des écoles sélectionnées, la prise de contact avec les cantons et les écoles, l'envoi du matériel et des instructions de test, la réception du matériel de test rempli par les écoles et la répartition des cahiers de test remplis entre les consortiums. La réalisation du test est décrite de façon détaillée dans un rapport interne (Mangold, 2007).

Evaluation: le partage des tâches entre les consortiums et le secrétariat général de la CDIP a fait ses preuves. D'autres structures sont toutefois préférables pour des projets semblables qui ne seront pas utilisés comme base d'HarmoS, mais serviront au monitorage de l'éducation. L'archivage des données sera achevé en été 2008, mais il manque une institution pour se charger de la gestion professionnelle du matériel de test HarmoS, ce qui est considéré comme un inconvénient pour l'utilisation du matériel de test HarmoS dans l'optique de la mise en oeuvre des modèles de compétence dans les cantons.

La surveillance du passage des tests par les enseignants eux-mêmes n'a pas entraîné de problèmes qui seraient apparus au moment de l'étalonnage. Il faudra toutefois clarifier une nouvelle fois l'adéquation de ce type de contrôle pour d'ultérieures études de monitorage.

Saisie des données

Le Secrétariat général de la CDIP a remis les cahiers de test aux consortiums qui ont pris en charge la saisie des données. Ils ont créé des fichiers Excel communs ou en fonction de la région linguistique, contenant les réponses des élèves sous forme codée. Lorsque les réponses aux exercices étaient ouvertes, il fallait un codage qui repose sur des instructions précises pour garantir une interprétation uniforme des types de solution.

Le groupe Méthodologie et les consortiums ont convenu de réexaminer l'adéquation des codages en cas de questions ouvertes; en général, cette vérification était intégrée dans la phase de la formation des encodeurs. 30 réponses au minimum devaient être codées de façon indépendante par deux personnes et ce pour chaque exercice à réponse ouverte. Le groupe Méthodologie a mis, pour l'évaluation, un document Excel à disposition dans lequel les codes pouvaient être introduits et le taux de congruence mesuré (Kappa, pourcentage de la concordance). L'évaluation de ce contrôle d'adéquation était entre les mains des consortiums. Exemple: le kappa linéaire moyen était, pour 33 items choisis au hasard, de .92 (extensibilité .77 – 1.0 avec une valeur déviante de .64) et une concordance moyenne de 96.3%.

Evaluation: la saisie des données s'étant déroulée sous la régie des consortiums, le groupe Méthodologie ne peut pas donner davantage d'informations si ce n'est que cela semble s'être bien déroulé. En tout cas, les données étaient utilisables pour l'évaluation. Le codage uniforme des questions à

réponses ouvertes par delà les frontières linguistiques représente toutefois un véritable défi qui n'a pas toujours pu être relevé avec succès. Il faut savoir que le codage fiable des réponses aux questions ouvertes en trois langues (textes écrits) exige bien davantage de ressources que les consortiums n'en disposaient. Les problèmes des tests en trois langues (traduction, correction des réponses) ont eu des répercussions sur les données qui ont mis en évidence certaines faiblesses dans la traduction et dans la correction des tests. Les futures enquêtes de monitorage devront veiller davantage à garantir une unité interlinguistique que cela n'a été le cas pour l'étude de validation.

Préparation des données

Les consortiums ont produit plusieurs centaines de fichiers qui contiennent chacun les résultats d'un cahier de tests. Pour être évalués, ils ont dû être transcrits pour chaque discipline dans une matrice de données, dont chaque colonne représente un item et chaque ligne une personne testée. La conception du test a entraîné une majorité de cellules vides dans cette matrice (par ex. mathématiques, 9^{e} année à 92.5%). Cette transcription a été effectuée par Edi Böni sous la supervision d'Erich Ramseier.

Dans les sciences expérimentales et les langues étrangères, les réponses (par ex. séries vrai/faux) ont été saisies directement. Conformément aux indications des consortiums, elles ont été transformées en partial-credit-items avec un petit nombre d'échelons d'évaluation.

Evaluation: La préparation des données a consommé beaucoup de temps il est vrai, mais elle n'a pas posé de problèmes sérieux.

Pondération

En plus des échantillons internes aux écoles, Ramseier et Moreau (2007) décrivent la pondération et la procédure d'estimation nécessaires pour évaluer les échantillons complexes et estimer les erreurs d'échantillonnage.

Evaluation: même si la procédure a bien fonctionné, elle n'a été toutefois que d'une utilité modeste pour l'étude parce qu'il n'y a guère eu plus de personnes observées que la part pondérée dans les niveaux de compétence.

6. Etalonnage

Choix du modèle

L'étalonnage et l'évaluation des données du test ont été au coeur des activités du groupe Méthodologie. Le test HarmoS avait pour objectif de vérifier un grand nombre d'exercices avec leurs contextes pour tirer au clair leur aptitude à caractériser un modèle de compétences. Le degré de difficulté des exercices une fois connu, ces derniers ont pu servir à illustrer les niveaux de compétence pour aboutir ensuite à une description de la compétence axée sur des critères.

L'étalonnage s'est basé sur un modèle de Rasch (pour en savoir plus, consulter l'annexe du guide). Le modèle de Rasch part de l'existence d'une relation entre la compétence à mesurer, qui n'est pas directement observable, et la probabilité de résoudre un exercice. Ce lien est décrit par une fonction mathématique (courbe de la caractéristique de l'item), qui, indépendamment de la translation, devrait être identique pour tous les exercices[2].

Deux raisons ont été déterminantes pour l'utilisation de la théorie des réponses aux items:

1) A l'instar d'autres modèles de la théorie des réponses aux items, le modèle de Rasch peut placer les exercices et les individus sur une même échelle qui peut servir à mesurer les exigences ou les compétences. Les exercices sont classés en fonction d'une probabilité prédéfinie (par exemple 50%) avec laquelle des individus dotés d'une compétence également définie pourront résoudre cet exercice. L'emplacement des exercices permet de caractériser et de décrire le contenu d'un domaine de l'échelle des compétences. C'est ainsi que la relation matérielle avec le modèle de compétence théorique peut être vérifiée.

2) Le modèle de Rasch et d'autres modèles de la théorie des réponses aux items permettent d'inclure de nombreux exercices dans une analyse cohérente même si chaque exercice n'a été traité que par une petite minorité d'élèves.

Evaluation: La validation empirique de données ne peut se faire sans étalonner ces données par une procédure scientifique dont l'application présuppose diverses conditions qui influencent directement l'élaboration et surtout le format des exercices. En effet, tous les exercices utilisés à l'école pour l'apprentissage et les contrôles ne se laissent pas transformer en un exercice de test. Ensuite, de nombreux formats ont de fortes répercussions sur les coûts. Les personnes responsables ou concernées (dans la politique et l'administration de l'éducation, l'école, les parents, les élèves, le grand public) doivent impérativement tenir compte du fait que, d'une part, les formats choisis ne permettent de tester qu'une partie des compétences et que, d'autre part, pour des raisons financières, certaines compétences fondamentales comme la production langagière (élocution et écriture) sont souvent laissées de côté. On a essayé autant que faire se peut de saisir toutes les compétences dans le cadre de la validation, y compris les compétences productives (élocution et écriture). Néanmoins les exercices de test validés empiriquement jusqu'ici ne recouvrent qu'une partie des compétences définies par HarmoS.

[2] D'autres modèles de la théorie des réponses aux items permettant des variations supplémentaires de cette fonction (cf. annexe du guide) auraient également pu être utilisés. Préférence a été donnée au modèle de Rasch car, outre sa simplicité conceptuelle, il rend possible des estimations plus stables pour un nombre d'individus relativement réduit.

Sélection des items

Ajustement avec le modèle de Rasch

Tout d'abord, on a vérifié si les exercices étaient appropriés, opération d'autant plus importante qu'aucun prétest complet n'avait été effectué. Le premier critère de cet examen était l'adéquation des items au modèle de Rasch, c'est-à-dire une correspondance satisfaisante de la relation empirique entre compétence et probabilité de réponse à la fonction postulée sur le plan théorique. L'outil utilisé, indice d'ajustement interne *infit* (moindre carré pondéré ou weighted mean square), montre pour chaque item combien de réponses inattendues sont observées en partant du modèle de Rasch. L'*infit* a une valeur attendue de 1. Une valeur fit trop élevée indique que la sélectivité des items est trop basse et, inversement, une valeur fit trop basse indique une sélectivité trop élevée. Le choix des items suivait un *infit* qui ne devait être ni inférieur à 0,70 ni supérieur à 1,30 (Wright & Linacre, 1994).

Evaluation: Le critère de l'*infit* n'a pas été critique et a été satisfait par la majorité des exercices. Pour les mathématiques, 9e année, il n'a fallu exclure qu'un exercice sur 269. Le choix du modèle de Rasch comme base de l'étalonnage n'a donc eu pratiquement aucun effet restrictif.

Sélectivité

Les exercices devaient remplir un deuxième critère, la sélectivité. Par sélectivité, on entend la corrélation des items avec le résultat global des tests. La fourchette des valeurs se situe entre –1 et +1. La sélectivité des items ne devrait en aucune manière être inférieure à 0.30. Une faible corrélation peut tenir à une formulation ambiguë de l'exercice ou au fait qu'une approche erronée peut tout de même aboutir au résultat correct ou encore que des éléments extérieurs facilitent ou compliquent la solution, etc.

On a donc suffisamment de bonnes raisons pour exclure les items à sélectivité insuffisante ($r < 0.3$), mais en la tolérant pour les exercices particulièrement faciles ou difficiles. Même conformes au modèle de Rasch, leur sélectivité[3] est basse: quand (presque) tout le monde peut les résoudre, il ne peut (presque) pas y avoir de relation entre la fréquence de la solution et la compétence au sein de cette population particulière. On peut conserver l'item pour démonter ce dont tout le monde est capable. Cela peut avoir du sens quand il ne s'agit pas d'un test portant sur l'efficacité mais illustrant l'envergure d'une compétence. Un vaste champ d'exercices très simples est toutefois envisageable et la vérification empirique ne peut pas démontrer une relation entre le type de réponse (presque) toujours juste à ces exercices et la compétence concernée.

Evaluation: dans l'ensemble, il faut exclure nettement plus d'exercices en considération d'un manque de sélectivité plutôt que d'un infit insuffisant. Les défections ne sont pas excessives vu l'absence du prétest (cf. documents de chaque consortium).

[3] On utilise le terme de «sélectivité» dans des acceptions très diverses. D'une part, dans la théorie classique des tests, elle est définie comme la corrélation entre l'item et le test (par ex. Krauth, 1995, p. 266); d'autre part, dans la théorie des réponses aux items, elle représente la courbure maximale de la caractéristique de l'item (par ex. Rost, 2004, p. 98). Dans le modèle de Rasch, tous les items ont la même courbure maximale de la caractéristique de l'item; leur sélectivité classique varie selon la population et elle est faible quand l'item est excessivement facile ou difficile pour cette population.

Différences linguistiques

L'invariance de la mesure est le troisième critère. Cela signifie que le test utilisé pour différents groupes de personnes mesure toujours la même chose – en d'autres termes, le modèle de compétence décrit la même chose.[4] Ce n'est pas le cas par ex. lorsque l'exercice 1 dans le groupe A est particulièrement difficile, en revanche très facile dans le groupe B, tandis que c'est exactement l'inverse pour l'exercice 2. Ces différences peuvent être dues à toutes sortes de causes externes. Il est néanmoins possible que les groupes A et B se soient structurés de façon différente ce qui serait plausible dans la mesure où les compétences sont des aptitudes *apprises* et que les possibilités d'apprentissage peuvent varier d'un groupe à l'autre.

Quand on parle de *groupes* chez HarmoS, les régions linguistiques sont au premier plan: les opportunités d'apprentissage peuvent différer culturellement d'une région linguistique à l'autre, l'uniformité par delà les langues joue un rôle clé dans un projet national dans la Suisse plurilingue. Des problèmes de traduction peuvent en outre être à l'origine du comportement variable des items dans les régions linguistiques.

Le critère de l'invariance linguistique (fonctionnement différentiel de l'item) s'est avéré particulièrement épineux. En effet, une part assez considérable des items fonctionne différemment d'une langue à l'autre. Le groupe Méthodologie a par conséquent opté dans le choix des exercices pour un critère généreux et pragmatique: un item, existant en français, allemand et italien était considéré d'égale difficulté quand les trois paramètres calculés séparément ne dépassaient pas 1 Logit ou quand aucun des items ne déviait de plus de 0.5 Logit de la valeur moyenne établie en commun. Quand les items n'existaient qu'en deux langues, le critère était adapté conformément à la plus faible probabilité de déviance qui est de 0.816 Logit, en l'occurrence à 0.408 Logit.

Afin de pouvoir les utiliser même lorsqu'ils ne remplissaient pas ce critère, certains items ont été «régionalisés», c'est-à-dire que l'item est traité pour chaque région comme un item autonome qui n'est pris en compte que dans cette région. Il n'a donc plus d'influence sur la comparaison entre régions tout en pouvant servir à décrire une compétence, les particularités de cette région étant prises en compte.

Evaluation: Nous revenons sur le fait que la charge considérable représentée par le développement de tests en trois langues a été nettement sous-estimée. A l'avenir, il faudrait en tenir compte davantage. Les critères appliqués pour la mise en évidence des différences linguistiques sont très souples, bien moins stricts que les critères de développement des tests pour la comparaison internationale des résultats scolaires PISA. C'est une situation qui était également insatisfaisante pour le groupe Méthodologie. L'objectif spécifique de cette étude et le temps limité disponible ont incité à utiliser un critère très généreux et à légitimer la sélection des items avec pragmatisme. L'objectif était que les différences linguistiques ne fassent pas perdre trop d'items. En effet, une trop grande rigueur dans la procédure aurait signifié l'échec de la validation nationale des modèles de compétence. Dans ce cas, le projet aurait dû se poursuivre au niveau des régions linguistiques.

[4] Par ex. Embretson et Reise (2000, p. 250).

Dimensionnalité

Les modèles théoriques postulent des éléments et des domaines très différents de compétence. On peut ainsi vérifier si aux dimensions théoriques correspondent des sous-échelles pouvant se distinguer l'une de l'autre tout en ayant une fonction de corrélation leur permettant d'être considérées comme une composante d'une compétence. Utilisé pour étalonner, le programme «ConQuest» (Wu et Adams, 1998) permet d'évaluer et de comparer des modèles pluridimensionnels avec un modèle unidimensionnel. «ConQuest» donne pour chaque modèle une valeur d'écart statistique par rapport aux données, appelée déviance et à interpréter comme un chiffre relatif. Cette dernière ne peut par conséquent être comparée avec la déviance d'un autre modèle que lorsqu'on utilise à chaque fois le même ensemble de données, c'est-à-dire le même nombre d'items etc. En règle générale, plus un modèle est différencié (nombre de dimensions), plus la déviance tend à se réduire. Lorsqu'un modèle ayant de nombreuses dimensions fait l'objet d'une vérification, il y a de fortes probabilités pour que la déviance de ce modèle pluridimensionnel soit plus faible que celle d'un modèle unidimensionnel. Pour vérifier que tel modèle pluridimensionnel avait une adéquation nettement meilleure avec les données qu'un modèle unidimensionnel ou qu'un autre modèle pluridimensionnel, les déviations ont été évaluées au moyen d'un test du Chi 2 et de mesures informationnelles (CAIC et BIC). Une étude de simulation d'Antonietti et Moreau (2008) indique que cette vérification n'est qu'une protection insuffisante contre une éventuelle meilleure adéquation d'un modèle pluridimensionnel doté de plus de degrés de liberté.

Evaluation: Les résultats varient selon le consortium. Il en ressort toutefois que les modèles créés sur la base de dimensions postulées sur le plan théorique correspondent mieux aux données qu'un modèle unidimensionnel. Cela ne veut cependant pas dire que des structures radicalement différentes ne conviendraient pas encore mieux. L'attribution de certaines tâches à des sous-dimensions n'est pas vérifiée. Le manque de temps n'a pas permis d'approfondir la dimensionnalité en comparant différents modèles entre eux.

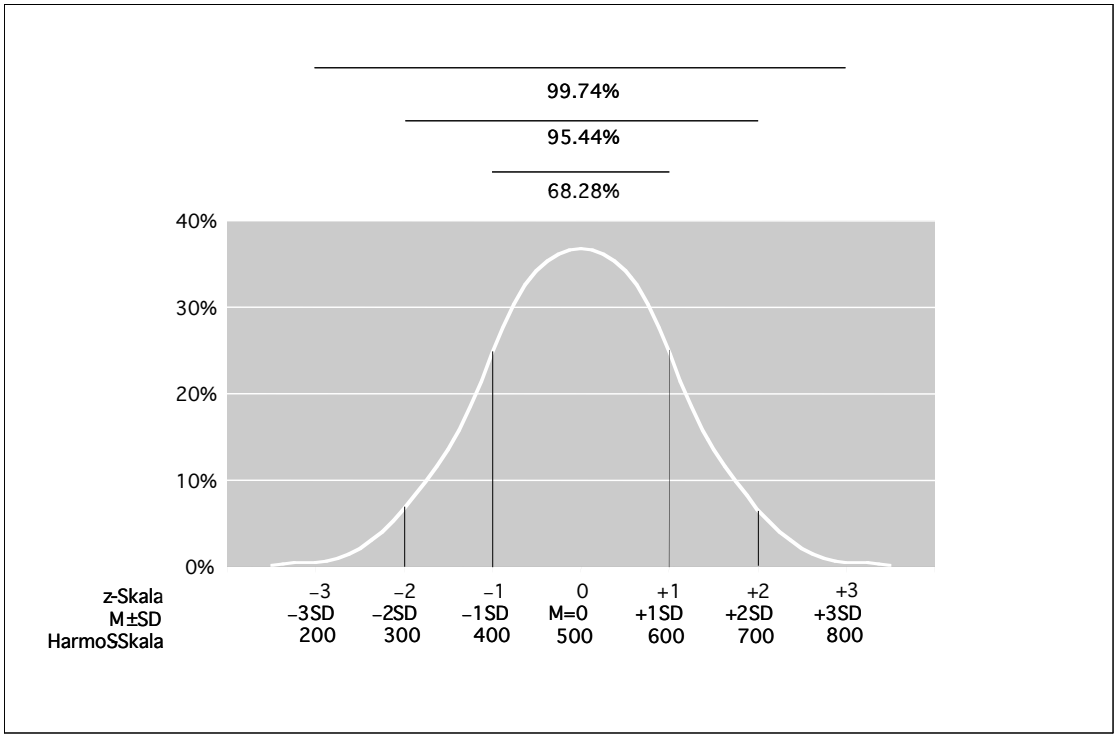
Transformation de l'échelle

La relation entre probabilité de réponse, difficulté de l'exercice et capacité individuelle dans le modèle de Rasch part d'une probabilité de réponse p = 0.5 toutes choses égales par ailleurs. Si maintenant il faut définir les capacités d'un élève sur la base d'exercices dont la difficulté correspond à ses capacités, cela veut dire que l'élève a une probabilité de réponse de 50%. Cette probabilité est trop faible pour permettre de décrire des standards de formation: une probabilité de 50:50 n'exprime pas que l'élève maîtrisera l'exercice à coup sûr. Voilà pourquoi le groupe Méthodologie a décidé de transformer les indices de difficulté de manière telle que la probabilité de réponse s'élève à p = 0.67 (2/3). Cette transformation déplace tous les items vers une compétence plus élevée. Les exercices qui sont au seuil de réalisation d'un standard de formation seront résolus avec une probabilité de 67% par les personnes qui satisfont tout juste à ce standard.

Dans le cas d'exercice dichotomique, cette transformation est un déplacement des paramètres des items de même valeur; elle ne change que faiblement leur interprétation. Lorsqu'il s'agit d'exercices qui reçoivent plusieurs points selon la solution (X = 0,1,2,3…), on recourt à des paramètres de seuil clairement

compréhensibles (par ex. quand p(X ≥ 2)) au lieu des paramètres directs du modèle. Cette transformation des paramètres de seuil a été dérivée par J.-Ph. Antonietti des paramètres du modèle au moyen d'un algorithme aboutissant à certains déplacements sous les seuils selon la configuration globale des items.

En appliquant le modèle de Rasch, les indices de difficulté et les paramètres individuels varient autour de zéro et les résultats sont généralement disponibles sous forme de chiffres allant de –3 à +3. Les valeurs négatives ne sont pas un moyen solide pour décrire la compétence d'une personne. Pour empêcher chiffres négatifs et décimales, le groupe Méthodologie s'est décidé à procéder à une deuxième transformation. On a transformé linéairement ensuite les indices de difficulté et les paramètres individuels de façon que la moyenne de la nouvelle distribution s'élève à 500 et l'écart-type à 100 points.



Une transformation linéaire de ce type ne change rien au contenu des expressions. En revanche, de même qu'une transformation z en une valeur moyenne de zéro et un écart-type de 1, elle facilite l'estimation d'une valeur individuelle par rapport à la valeur moyenne et aux déviations habituelles, ce qui favorise des expressions tendant vers la norme. Ceci est d'autant plus vrai que les valeurs tests rapprochées ont une distribution normale.[5]

Environ 68 pourcents des résultats des tests se situent dans une fourchette entre la valeur moyenne ± un écart-type, tout juste 95 pourcents entre la valeur moyenne ± deux écarts-type, soit entre 300 et 700 points. On a opté pour l'échelle dont la moyenne est égale à 500 et l'écart-type à 100 en accord avec PISA.

[5] La forme exacte dépend du modèle d'échelle utilisé. ConQuest part d'une distribution normale de la distribution latente; quand des caractéristiques complémentaires (âge, type scolaire) interviennent dans l'estimation, l'hypothèse d'une distribution normale pour les valeurs résiduelles correspondantes s'impose. Concernant la logique suivie dans l'estimation et les transformations, se référer au rapport technique PISA, OCDE (2003).

Appréciation globale de l'étalonnage: l'utilisation du modèle de Rasch a fait ses preuves, tout en n'entraînant pas de limitations exagérées au sein des exercices repris dans le test ni du contenu qu'ils représentent. D'entrée de jeu, les conditions d'une évaluation à grande échelle (large-scale-assessements) défendable sur le plan économique ont fixé certaines limites. Le recours au modèle de Rasch ou à un autre modèle de la théorie des réponses aux items est nécessaire pour rendre possible l'évaluation des données en dépit de la distribution sur un nombre de personnes relativement petit. Cela permet de classer les exercices et les individus sur la même échelle pour pouvoir interpréter les tranches de l'échelle de compétences sur la base du contenu des exercices.

D'une région linguistique à l'autre, les exercices étaient de difficulté variable, ce qui a posé problème. C'est un sujet à creuser sur la base des données disponibles pour pouvoir distinguer les problèmes de traduction, la différence de culture et de tradition d'apprentissage. Les futures enquêtes devront veiller à faire concorder le choix des exercices avec la représentativité régionale, si l'on ne veut pas limiter le contenu au plus petit dénominateur commun entre les régions linguistiques.

Pendant le laps de temps consenti, seuls les critères essentiels de l'accomplissement du modèle et de l'ajustement des items ont pu être vérifiés. Le contrôle de la dimensionnalité est particulièrement lacunaire. D'autres critères devraient être passés au crible avant l'utilisation ultérieure des exercices (voir ci-dessous).

7. Définition des niveaux de compétence et des standards

L'étalonnage est au premier chef une dimension continue des compétences et des exigences dans laquelle on classe des personnes (avec leurs compétences) et des exercices (avec leurs exigences). Quant au modèle de compétence, il parle de niveaux et de gradation sur cette dimension. On peut alors se demander comment définir et décrire des niveaux discrets et quelle est leur relation avec l'échelle continue primaire de l'étude de validation. Deux approches opposées sont envisageables: priorité à l'échelle et interprétation a posteriori (démarche suivie apparemment en L1) ou formulation a priori des niveaux de compétence et recherche d'une relation avec l'échelle (tendance plutôt M). Il existe certes des modèles sur la manière d'utiliser des caractéristiques spécifiques d'exercices définies a priori, pour marquer des niveaux d'exigence; ce n'est pour l'instant qu'une approche qui présuppose une classification exhaustive des exercices, ce qui n'était pas réalisable, toujours pour des questions de temps disponible. La relation entre conception théorique et échelle a dû bien plus être créée par des procédures d'interprétation non formalisées.

Si l'on part en premier lieu de l'échelle continue, les intervalles peuvent être plus ou moins larges. En général, ils s'élèvent à une unité (logit). L'avantage de cette procédure est la similitude relative de l'interprétation des résultats: si un élève est affecté à un niveau, la règle générale veut qu'il puisse résoudre en moyenne au moins 50 pourcents des exercices de ce niveau. Si l'on suit plutôt les caractéristiques théoriques, on peut s'attendre à une variation des intervalles.

Toujours en raison du temps, les consortiums ont défini une majorité des niveaux sans consulter le groupe Méthodologie. Ce faisant, ils se sont inspirés des objectifs cibles formulés dans le guide, qui étaient de 4 niveaux. Par la suite, la définition des niveaux varie fortement (notamment dans la largeur sur l'échelle).

Les standards de base ont eux aussi été en grande partie déterminés par les consortiums sans collaboration étroite avec le groupe Méthodologie. Ce dernier avait cependant présenté les méthodes courantes permettant de systématiser la définition d'une valeur limite. Leur utilisation ne changerait rien au fait que des standards sont toujours de l'ordre de la prédétermination et de l'estimation. Ils permettraient néanmoins de rendre le processus de détermination transparent et de systématiser l'inclusion de différents groupes d'individus (acteurs ou stakeholders).

L'étude de validation a permis de définir la part (variable selon les consortiums) des élèves qui remplissent les standards proposés (voir leurs rapports respectifs). Le guide postule clairement à la p. 9: «les standards de base doivent en premier lieu être justifiés sur le plan du contenu. Ils doivent toutefois être réalistes.» Si l'on en croit une publication de de Pietro, Müller et Wirthner (2008, p. 44, 46), on a au contraire appliqué pour la langue première une orientation sur la norme statistique, reposant même sur une indication par la CDIP de 20%.

Evaluation: les consortiums ont défini en grande partie de façon autonome les niveaux de compétence et les standards de base dans l'urgence. Le groupe Méthodologie ne peut donc formuler que peu de commentaires à ce sujet. L'utilisation, même si elle était coûteuse, des méthodes formelles connues a permis de mieux conforter les niveaux et les standards. Pour une dimension continue primaire des compétences, les niveaux discrets ne sont que des

compléments à la caractérisation des domaines de compétence sans marquer de réels degrés.

La procédure semble avoir varié d'un consortium à l'autre et parfois contredire la conception originelle. En raison de l'importance conférée aux niveaux et aux standards (au contraire de la compétence continue primaire validée), la méthodologie et la transparence des définitions ne sont pas satisfaisantes. La transparence de la communication, et plus particulièrement dans le contexte scientifique, est tout aussi essentielle que la prise de conscience par les milieux politiques de l'expressivité limitée des standards. Vu l'enjeu de l'étude, les consortiums devraient avoir à disposition une année de plus pour approfondir la définition des standards avec des représentants des milieux scientifiques et professionnels ainsi qu'avec d'autres acteurs. La manière pratiquée jusqu'ici de fixer les standards ne semble pas avoir suffisamment tenu compte des règles d'une procédure explicite, fondée sur des bases scientifiques.

8. Archivage

Lorsque le travail avec les consortiums sera terminé, les données utilisées pour l'évaluation seront réunies sur un DVD et transmises à la CDIP. Elles restent ainsi disponibles pour d'autres évaluations ou en guise de lien avec les prochaines enquêtes de monitorage.

9. Conclusions

Les évaluations comportant ou soumettant des conclusions ont déjà été insérées dans les différentes sections du rapport. Voici quelques considérations générales.

Etat de l'étude de validation

L'étude de validation a porté ses fruits. Le simple fait que dès le début les modèles de compétence aient pu être illustrés par des exercices dont la difficulté effective n'est pas seulement présumée mais connue, justifie la démarche. Il faut toutefois constater que la validation est un processus de longue haleine dont seul une première étape a pu être franchie dans le cadre existant. En outre, au vu des 10 000 participants et du nombre considérable d'exercices de test, l'étude de validation est une entreprise de taille dans la recherche éducationnelle empirique en Suisse, qui néanmoins n'a été évaluée que de façon limitée. Du point de vue du groupe Méthodologie, une base pour la validation scientifique des modèles de compétence a été créée plutôt qu'une véritable possibilité de déjà valider les modèles. Un bon départ a été pris, mais l'objectif n'est pas encore atteint dans la mesure où en Suisse les modèles de compétence doivent encore être validés empiriquement au sens scientifique du terme et où les standards restent à définir selon une procédure scientifique.

Suite des opérations

Le processus de validation devrait être poursuivi en vue d'un futur monitorage valide de la scolarité obligatoire. Certaines questions de recherche devraient être définies dans une optique mixte à la fois spécifique et méthodologique. D'autres mériteraient d'être mentionnées pour des considérations méthodologiques. Les thèmes dans la liste ci-dessous sont triés selon qu'ils peuvent être examinés en fonction de l'évaluation des données existantes (suite de l'évaluation) ou qu'ils exigent de nouvelles collectes de données dans un cadre plus restreint (travaux de suivi).

Evaluations ultérieures

a) Comment se manifestent les différences de cursus dans les résultats des tests? Par l'analyse des échelles, sous-échelles et items individuels (fonctionnement différencié des items ou differential item functioning DIF) selon le type scolaire et parfois le canton

b) Comment se manifeste le fait d'être issu de la migration ou allophone dans les résultats des tests (DIF, comparaison entre sous-échelles)?

c) Réaliser d'autres tests type, par ex. DIF en fonction des élèves forts/peu doués

d) Déterminer plus complètement les erreurs des valeurs d'estimation utilisées; et aussi mieux adapter aux indications théoriques la définition empirique de la partie de population considérée (ex. mathématiques: calcul actuel sur la base de l'échelle totale; oralement, on demande la réalisation des standards pour chaque sous-échelle)

e) Examiner la dimensionnalité: attribution multiple des exercices à des échelles (contrairement à l'analyse précédente, plusieurs exercices peuvent toucher plusieurs éléments de la compétence); analyse dimensionnelle exploratoire

f) Mettre en relation les compétences entre les différentes disciplines; par ex. analyse de la compétence en mathématiques en contrôlant la compétence en lecture. Quelle part des élèves échoue dans les standards de formation dans toutes les quatre disciplines ou dans une discipline au moins?

g) Prédire la difficulté des items à partir d'autres caractéristiques des exercices: préciser davantage la description des niveaux de compétence

h) Arriver à saisir la difficulté des exercices est le résultat central de la validation. Cette dernière a pu être définie malgré les limitations inhérentes aux conditions dans lesquelles s'est déroulée l'étude. Par exemple: les exercices sont regroupés dans un cahier de test qui les présente dans un ordre particulier. L'évaluation montre que dans certains cahiers de test, les derniers exercices n'ont souvent pas été résolus. Ces informations manquantes n'ont pas été prises en compte dans l'appréciation de la difficulté des exercices. Il est toutefois probable que les élèves plus faibles ont sauté des exercices parce qu'ils ont été très souvent confrontés au manque de temps. La difficulté des exercices a dû de ce fait être sous-estimée. La portée de cet effet se laisse analyser:il existe un lien entre la fréquence de l'omission des exercices, les capacités des personnes concernées et le rang de l'exercice.

i) La motivation est un élément de la compétence. Les questionnaires pour les élèves contenaient des indications quant à la motivation; il aurait fallu mieux les évaluer et les mettre en relation avec les modèles de compétence.

Travaux de suivi

a) Une étude complémentaire avec de nouveaux échantillons et une autre séquence d'exercices permettra d'approfondir la portée des indications manquantes.

b) Une suite fixe d'exercices dans les cahiers de test a d'autres conséquences qu'une accumulation d'exercices non résolus à la fin des cahiers. Si l'on veut créer un contexte de travail logique dans les cahiers de test, de nombreux exercices doivent se référer à des contenus communs ou voisins. Il s'en suit une relation de dépendance entre les exercices, par ex. en résoudre un permet d'apprendre quelque chose d'utile pour l'exercice suivant. De telles relations de dépendance sont toutefois en contradiction avec les conditions de départ du modèle d'étalonnage utilisé. La probabilité de ces relations est confortée par le fait que le même exercice de mathématiques figurant par erreur dans deux cahiers de test n'a pas la même fréquence de résolution correcte. Là aussi, une étude complémentaire sur des exercices placés dans un ordre différent pourrait apporter des éclaircissements.

c) Certains exercices et thèmes ont été exclus de la validation quand ils ne pouvaient pas être appliqués concrètement dans le cadre donné. La validation pourra être étendue à ce type de questions dans de nouvelles études, ce qui permettra de lutter contre la menace que des standards et des tests restreignent la matière enseignée de manière inappropriée. Le consortium sciences expérimentales a inclus ce type de tests authentiques dans la première phase de validation.

d) Analyse des processus de développement par delà le degré scolaire.

Tous ces travaux jouent un rôle pour que l'on sache clairement quels exercices et comment peuvent être utilisés comme unité indépendante dans une future étude de monitorage, afin de créer un lien avec les standards définis maintenant. Ces travaux aboutissent également à des conclusions générales qui seront nécessaires pour la mise sur pied du monitorage. Ils présentent l'avantage pratique de pouvoir être réalisés avec des échantillons relativement petits qui tout en devant couvrir tout le spectre des résultats ne sont pas obligés d'être représentatifs.

Cadre organisationnel

Réalisée dans un cadre fortement limité, l'étude de validation a été également caractérisée par des omissions notables. Parmi les limitations, citons l'absence de prétests, le calendrier trop serré pour presque toutes les étapes du travail, l'activité du groupe Méthodologie exercée à titre d'activité professionnelle secondaire, l'interaction limitée et l'absence d'intégration entre domaines spécialisés et méthodologie.

Même si, grâce à l'engagement sans faille de nombreux acteurs, l'étude de validation a été partiellement couronnée de succès, il serait erroné de la reprendre telle quelle dans la planification des futures études de monitorage de l'éducation car elle perpétuerait ainsi cette situation marquée par la carence.

Monitorage de l'éducation et rapport entre didactique des disciplines et psychométrie

Il existait une nette séparation et une forte hiérarchisation dans l'étude de validation entre le travail sur la discipline et son contenu d'une part et le travail méthodologique d'autre part: la responsabilité était dévolue aux consortiums, le groupe Méthodologie quant à lui, par l'intermédiaire du groupe stratégique, donnant certes des directives, mais finalement exerçant surtout une double fonction de conseil et d'exécution.

D'une manière générale, cette distinction a entraîné que l'on recoure seulement partiellement aux résultats empiriques de l'évaluation et de l'interprétation du contenu et que l'on néglige les procédures existantes pour définir les niveaux et les standards.

La situation de l'étude de validation était particulière: dans un premier temps, il y allait de l'élaboration d'une construction théorique, le modèle de compétence; les résultats empiriques n'avaient qu'une fonction d'appui. En revanche, les éléments concernant le contenu disciplinaire bénéficient à juste titre d'un rôle privilégié. Il en ira autrement dans les futures études de monitorage: en se fondant sur les modèles de compétence existants, on veut saisir de façon fiable

la distribution de la compétence sur les régions et les cantons et saisir les caractéristiques des écoles et des individus, qui viennent appuyer l'interprétation des différences. A cela s'ajoute de devoir garantir que des études postérieures pourront dégager des tendances avec certitude. Tout cela pose de grandes exigences à la méthodologie.

Bien entendu, les éléments de contenu disciplinaire joueront également un grand rôle dans les futures études dans le contexte du monitorage de l'éducation. Priorité doit toutefois être accordée à une vision méthodologique analytique globale, ne serait-ce que pour donner un cadre uniforme des différentes disciplines pour le monitorage.

Monitorage de l'éducation et structures de recherche

La recherche suisse en éducation pâtit du fait qu'à court terme sur un thème particulier, on lance des initiatives et des projets et on édifie des structures et des compétences. Une fois les projets terminés, ces structures avec leurs compétences se désagrègent. Or elles font cruellement défaut lorsque le thème redevient d'actualité.

Et c'est justement la menace qui pèse sur l'après HarmoS, puisqu'il est prévisible que les compétences créées au sein d'HarmoS seront bientôt nécessaires pour le monitorage de l'éducation. Soutenir des axes de développement en relation avec des groupes de disciplines permettrait de contrer cette menace. En raison de la fragilité des bases existant en Suisse, il faudrait surtout promouvoir et asseoir sur le plan institutionnel l'expertise méthodologique et la mesure de compétence. Nos voisins allemands et autrichiens ont réagi à la situation de carence existant chez eux en créant des structures et parfois en lançant des programmes de recherche. A la différence de ces pays, la Suisse a d'entrée de jeu fondé le développement de standards de la formation sur des bases empiriques. Cette œuvre de pionnier doit maintenant prendre des formes durables si la Suisse ne veut pas perdre son avance gagnée grâce à une démarche innovante.

Perspectives

La réussite de la mise en œuvre des standards de la formation est subordonnée à divers facteurs que nous évoquerons brièvement ci-dessous:

- Les résultats obtenus jusqu'à présent permettront à l'avenir d'orienter l'assurance qualité en comparant les résultats mesurés aux modèles de compétence fondés sur la didactique des disciplines en partie validée empiriquement.

- Les données existantes permettent de poursuivre la validation des modèles de compétence, d'ancrer le principe d'une validation empirique des modèles de compétence selon les règles de l'art et de conclure les travaux commencés en suivant une certaine logique.

- La mise en œuvre des standards de formation s'effectuera à deux niveaux. Premièrement, le prochain monitorage de l'éducation en vérifiera l'application rigoureuse. La CDIP a les outils nécessaires pour organiser les tests de monitorage de telle manière qu'ils correspondent aux modèles de compétences dont ils poursuivront également la validation.

- En second lieu, les standards de formation servent à faire un bilan de compétence de chaque élève. Le développement des bases nécessaires est entre les mains des conférences régionales linguistiques de la CDIP. La manière de résoudre ce problème n'est pas encore définie, mais il est très probable que certaines institutions mettront spontanément des tests sur le marché. Le fait que HarmoS ou la CDIP ne fournissent pas d'outils mais uniquement des modèles recèle le danger d'une dynamique incontrôlée dans le développement des tests. Au sens strict, les modèles de compétence ne sont pas décisifs dans l'harmonisation par les standards mais bien les échelles scientifiques qui permettent une mesure uniforme pour décrire et évaluer les résultats scolaires.

  Le monde politique serait bien inspiré de déterminer des règles claires sur l'utilisation du concept «Test HarmoS». Sans cela, il y a un risque de prolifération des étiquettes. Toute institution qui met un test sur le marché prétendra qu'il est HarmoS-compatible. Les solutions type Klassencockpit et Stellwerk seront alors vendues avec l'étiquette Test HarmoS. Si l'on veut être puriste, le concept «Test HarmoS» devrait être rattaché à une échelle nationale ce qui constituerait une protection. Si les concepts «Test HarmoS» ou «HarmoS-compatible» peuvent être utilisés sans être assortis de conditions (fixées empiriquement), l'ensemble du projet perd sa valeur sur le plan méthodologique. Ce scénario paraît vraisemblable au groupe Méthodologie parce qu'il permettrait d'économiser les coûts pour des tests HarmoS coûteux.

- Il serait judicieux d'assortir l'utilisation de l'étiquette «HarmoS-compatible» par les fournisseurs et réalisateurs de tests à la condition de mettre à disposition pour leur analyse scientifique leurs exercices et les résultats anonymisés des tests (données de bases en plus des valeurs d'échelle). Cela permettrait de contrôler la qualité du test et de garantir le développement méthodologique.

- Les tests pour le bilan de compétence individuel et pour le monitorage de l'éducation ne peuvent pas être considérés comme deux entités radicalement différentes: il doit chaque fois s'agir des mêmes modèles de compétence, des mêmes standards et échelles nationales, si l'on veut éviter le chaos dans le système de formation. Pour en être sûr, il faut qu'une partie des exercices utilisés soit commune. En outre, une bonne coordination entre les deux types d'application peut créer des synergies profitables.

Appréciation générale

Il reste à espérer que les travaux réalisés pour la validation des modèles de compétence seront évalués de façon réaliste et leurs résultats provisoires traités avec loyauté. Toutes les instances concernées, notamment le Secrétariat général de la CDIP, ont déployé des efforts considérables. Ce grand engagement a permis de grandes réalisations. Il est toutefois essentiel de poursuivre avec un fort engagement scientifique les travaux de consolidation et de validation des modèles de compétence et des standards de formation.

10. Annexe

Littérature

Antonietti, J.-P., & Moreau, J. (2008). *Perspectives méthodologiques dans le cadre d'un projet d'harmonisation de la scolarité en Suisse.* Paper presented at the conference of the Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE), Genève.

De Pietro, J.-F., Müller, R., & Wirthner, M. (2007). HarmoS-L1: Vers des standards de base pour la langue de scolarisation. – In Richtung standards de base im Bereich der langue de scolarisation. *Babylonia* (4), 40-52.

Embretson, S. E., & Reise, S. P. (2000). *item response theory*. Mahwah, NJ: Lawrence Erlbaum.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Psychologie Verlags Union.

Mangold; M. (2007). *Durchführung der empirischen Phase. Projekt HarmoS.* Interner Bericht. Bern: EDK

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4. ed., pp. 681-699). Westport, CT: American Council on Education.

Ramseier, E. & Moreau, J. (2007). *Stichprobendesign und Gewichtung in der HarmoS-étude de validation.* Arbeitspapier. Bern: EDK

Renaud, A. (2006). *Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles*. Neuchâtel: Office fédéral de la statistique.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest. Generalised item Response Modelling Software*. Melbourne: Australian Council for Educational Research.

HarmoS Test Design

The four consortia were responsible for developing the tasks and for most aspects of the test design like content coverage or choice of item formats, thereby supported by a guideline of the methodological group. They assigned each task to one cluster, each cluster taking 20 or 30 minutes of testing time. In total, they produced 209 clusters which had to be administered in German, French, and mostly in Italian (cf. table 1).

Since mathematics and science were tested in one session and languages in the other, these sessions could be designed separately. Both testing sessions consisted of two positions before and two after a break. In each position, students could work on one cluster. Since each student only worked on 8 clusters (matrix design), a main challenge was to distribute and combine clusters in a way that all tasks of a subject could be scaled on a common scale, correlations between content areas of competence and between aspects of competence could be estimated, and position effects were controlled at least between clusters.

Table 1: Number of Clusters/Booklets by Subject, Grade and Language Region

Subject	Grade 6				Grade 9			
	German	French	Italian	Total	German	French	Italian	Total
Duration: 20 minutes								
Mathematics	26	26	26	78	34	34	34	102
Science	24	24		48	24	24		48
First language	15	15	15	45	15	15	15	45
Foreign languages	8	8		16	16	16		32
Duration: 30 minutes								
First language	19	19	7	45	20	20	8	48
Foreign languages	4	4		8	8	8		16
Total	96	96	48	240	117	117	57	291
	Grade 6 and 9 together							
All Clusters	213	213	105	531				
Clusters with different content	209	209	105					

Since mathematics and science were tested in one session and languages in the other, these sessions could be designed separately. Both testing sessions consisted of two positions before and two after a break. In each position, students could work on one cluster. Since each student only worked on 8 clusters (matrix design), a main challenge was to distribute and combine clusters in a way that all tasks of a subject could be scaled on a common scale, correlations between content areas of competence and between aspects of competence could be estimated, and position effects were controlled at least between clusters.

The printing arrangement allowed for assigning booklets individually to persons and positions. Therefore, a balanced design was not created by assigning certain combinations of four clusters systematically to a limited number of test booklets (usual multi-matrix-design). Instead, each cluster was assigned to a single short booklet and these booklets were randomly assigned to persons and positions, thereby aiming for a number of principles. For mathematics, grade 9, where each cluster included tasks from one content area of competence, these principles are:

- Each cluster is tested equally often in total and in each position
- Each cluster is equally often combined with each cluster from other content areas in pairs of clusters
- Pairs of clusters are evenly distributed across the class and student sample

The assignment of the 34 mathematics clusters had to be combined with the assignment of the 24 science clusters. Pairs of mathematics and science clusters were combined randomly. To each cluster, the same number of students was assigned; therefore most students worked on a pair of clusters from each of the two subjects but some students worked on two pairs of mathematics clusters.

To distribute clusters to students and positions, a list of all possible cluster combinations which meet the mentioned design principles was constructed by programming it. Copies of this list were combined to create a series longer than the number of students. The cluster combinations within each list were randomly sorted. Finally, the series was matched to the stratified list of students.

As a result of this procedure, the mentioned principles are only attained with some minor random variation. Each student got a set of test materials with a specific combination of four booklets per session. If one only considers position 1 and 2 in the German part of Switzerland, already 791 different combinations of mathematics clusters entered analysis in grade 9 – each combination only 1 or 2 times. If one considers all 4 positions of the mathematics and science test, this augments to 2079 different cluster combinations – only 10 of them dealt with by two students. Given such diversity, possible interactions between clusters are optimally controlled and tasks are linked with each other in many different ways. Despite some random variation, the design is almost equal to a full balanced incomplete block design which is considered as unrealistically complex (Mazzeo, Lazer & Zieky, 2006, p. 685). This is helpful, since in the mathematics test of grade 9, 92.5% of all information is missing by design.

To generate the cluster combinations for the other three subjects, a similar approach was applied. For the languages, the situation was somewhat more complex: For first language some tasks refer to an acoustic stimulus presented to the whole class. The clusters of these tasks were administered at the beginning of the testing session and had to be assigned on class level. For foreign languages, clusters of English, German, and French as foreign language had to be handled. In total, about 100’000 individual links between clusters, positions and persons were produced.

Erich Ramseier, BiEv, ED Bern